Selecting and Organizing Data

A DATAshare Guidance Document

Iowa Department of Management
10/23/2013

Introduction

We all depend on data in today's world, and opening it up is, perhaps, the best way to realize its value and maximize its potential. Publishing data generated and held by State government agencies can expand its use beyond state government and encourage innovative ideas (e.g., web applications) that enhance the lives of our citizens. Re-usable public data can increase economic activity by generating new and rich content through new applications and services. It also increases the level of openness our constituents, who are increasingly data savvy, and have come to expect. Lastly, it is required by State Law:

- Taxpayer Transparency Act – Iowa Code Chapter 8G – requires state agencies to provide expected performance outcomes and past performance outcomes achieved with state funding
- Accountable Government Act – Iowa Code Chapter 8E – requires state agencies to provide the "widest possible" dissemination of performance measures and performance targets based on data used by the agency to evaluate its performance
- Examination of Public Records (Open Records) – Iowa Code Chapter 22 – requires government agencies to provide public access to records and data

This guide is intended to provide state employees information on:

- How to identify data to publish, and
- How to organize your data into meaningful datasets.

Contents

What data should we publish?

Choosing data to publish is your first step. You likely have a lot of data within your agency and it is not realistic to make all of it available right away. Getting started is often the hardest part. Creating a list can help. Attempt to identify reliable and accurate data created or maintained by or on behalf of your agency that:

- Is used for decision-making, and/or policy development
- Supports the Governor's or your own agency goals and objectives
- Improves the public's understanding of your agency and its operations
- Is or has been requested by citizens, the media, the Governor, and/or Legislators
- Helps you better understand key problems or issues your programs are trying to address
- You spend a great deal of time and money collecting and maintaining
- Facilitates collaboration with other state agencies
- You already post online as Excel spreadsheets or database extracts, or have coupled to web apps
- You have want to make more accessible
- You want more feedback on
- Your counterparts in other states are publishing

Once you have a list; prioritize it. There are a number of ways to prioritize your list – you could start with what you believe will be the most valuable, what would be the easiest to publish, or better yet, some combination of the two. The Agency Dataset Inventory template can help you organize this process. You may even want to consider reaching out to constituent groups to obtain feedback on your priorities. If the data is already online, consider reviewing your website analytics to see which ones are most popular. However, the key thing is to just get started - pick a dataset to start with (even if it is part of a larger dataset). Starting small may be what you need to do to learn and build momentum.

What is a dataset?

A dataset is a comprehensive collection of interrelated data from one data source (e.g. state vendor payments, state budget, crime statistics, unemployment data, active businesses, etc.).

What data should we include in our dataset?

When determining what data to include in your dataset, it is important to understand what users or potential users – whether they are citizens, businesses, legislators, developers, or other state agencies – need. Understanding the needs of the users will enable you to include data that is relevant.

Once you have determined the relevant data to include, you need to think about how the data can and should be summarized to make it more meaningful and understandable. While there

are some users who will want access to the "raw" data, most citizens will find visual summaries, such as a chart or map, far more helpful. In order to summarize data you will need to include data that allows you to categorize/group your data. For example, with vendor payment data the following are a few columns of data provided to help summarize the information:

- Name of department, institution or sub-organization making the payment
- Source of funds for the payment
- The legislative authorization (appropriation) for the payment
- Expense category assigned to the payment
- Name of vendor receiving the payment

By providing these columns of data in the dataset, citizens are allowed to summarize total payments by department, total payments by vendor, and total payments by fund, etc.

What data should we avoid publishing in datasets?

When considering data to publish, your focus should be on non-personal, non-confidential data. You should exclude data or take steps to remove data that:

- Contains information readily identifiable to a specific individual. Readily identifiable information would include, but is not limited to, such things as an individual's name, social security number, address, phone number, and email.
- Is deemed confidential or sensitive, or is protected by state (Iowa Code Section 22.7 or other applicable Iowa Code section) or federal law.
- Reflects the internal deliberative process of your agency or the State of Iowa, including but not limited to negotiating positions, future procurements, or pending or reasonably anticipated legal or administrative proceedings.
- Is subject to copyright, patent, trademark, confidentiality agreements or trade secret protection.
- Consists of employment records, internal employee-related directories or lists, and data related to internal agency administration

If excluding or removing the data creates an undue financial or administrative burden, consider how the data might be meaningfully aggregated before being made available to the public.

Should I include totals and subtotals in my dataset?

No, you shouldn't include rows or records in your dataset that are totals and subtotals. It is best to only store the raw data and leave totaling and subtotaling up to the system. Excluding

rows that total and subtotal helps ensure that calculations are always reflecting current values when the data is updated.

How detailed should our dataset be?

Granular (or detailed) data provides more options for presenting or summarizing the data. By providing a higher level of detail in your dataset, you facilitate making the data more meaningful and understandable to a wider cross section of the public. Because of this, you should always strive to make your dataset as granular as possible. However, there may be circumstances that require you to publish a dataset comprised of records that summarize underlying data. This usually occurs where record level detail is comprised of confidential or sensitive information that must be summarized before it is released to the public. Individual student test scores would be an example of this. In this case, the dataset could present test results by grade, school and district.

What format does my data need to be in?

Data is currently imported into DATAshare using CSV file format, which stores tabular data (both numbers and text) in plain-text form, see figure 1. This file format was selected because so many programs and applications support some variation of CSV for exporting, which makes

```
Function,Special Department,Department Number,Department,Count Date,Full-Time
Employees,Part-Time Employees,Temporary Employees,Male,Female,White,African
American,Latino,Asian/Pacific Islander,Native American/Alaskan Native,Undisclosed
Race,Disabled
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,12/23/2010,364,8,4,217,147,319,16,5,15,3,6,27
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,1/20/2011,363,7,4,217,146,317,16,5,16,3,6,27
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,2/17/2011,361,7,4,219,142,316,16,5,15,3,6,27
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,3/31/2011,361,7,4,219,142,317,15,5,15,3,6,27
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,4/28/2011,361,6,4,219,142,317,15,5,15,3,6,27
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,5/26/2011,357,6,4,216,141,315,15,5,14,2,6,26
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,6/23/2011,357,7,2,216,141,315,15,5,14,2,6,26
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,7/21/2011,354,7,2,214,140,312,15,5,14,2,6,26
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,8/18/2011,356,6,2,215,141,315,14,5,14,2,6,26
Administration and Regulation,"Administrative Services, Department of",005,Administrative
Services,9/29/2011,355,6,2,214,141,315,12,5,15,2,6,25
```

Figure 1. Example CSV file. One record per line, text is wrapped.

moving tabular data between programs with different and incompatible formats possible. Here are some basic characteristics of CSV files:

- There is typically one record per line (hard return after the record)

- Records divided into fields that are separated by delimiters (e.g. commas, semicolons, tabs). It is the individual fields within a record that become the dataset's columns when imported.
- Each record contained in the file should have an identical list of fields.
- The first row in the file should serve as your column titles or headings.
- Fields containing a line-break, double-quote, and/or commas should be quoted so the file can be processed correctly[1].

How do I include relational data?

CSV files do not support a relational or hierarchical structure, as every record within a CSV file is expected to have the same structure. Where such a structure exist within the application or program your data is stored, you will likely need to join information through a query from two or more tables based on a common field or fields (usually a code or ID) within your program or application before you export your data as a CSV file.

The characteristics of the join will be dependent on your data, but in most cases, you will likely want all records from the table containing your main data (i.e. the numeric data you want the public to summarize) and only those records from related tables where common field match, see example in figure 2 below.

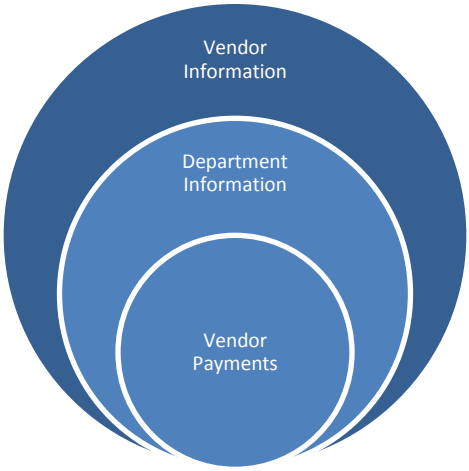

Figure 2. Capture all vendor payments, and only department and vendor information related to the vendor payments.

As an example, the vendor payments table (main data, shown in figure 2) stores only data about the payment itself, such as date, purpose, and amount. Rather than capturing all the information about a department or vendor, it only stores a department ID and vendor ID. The tables for the department information and vendor information contain additional information, such as names, addresses, etc. The department and vendor information tables could then be related (and joined to) the vendor payments using the IDs. This organization uses one record for each vendor and department, which can be related to many vendor payments. Once the tables are joined, the a dataset could include the department name, vendor name, and vendor address along with date, purpose and amount for each

[1] If you are saving an Excel file as CSV, Excel will automatically put quotes around text fields requiring them.

payment record. Any department information records or vendor information records not related to vendor payments would be ignored and not included in the dataset.

Take care to ensure your join does not duplicate numeric data you intend the public to summarize. This usually happens where the common field used in the join does not contain unique values.

Should I stack my data?

When working with data, it can be organized in a couple of different ways: multiple data columns (see figure 3) or stacked data column with category column (see figure 4). However, as you may note, multiple data columns does not mean the data is left unstacked. All datasets need at least one means of categorizing data in order to create most visualization types in DATAshare. In the example below, date serves as the means of categorization. Both multiple data columns and stacked data can work for publishing and presenting your data in DATAshare. As you begin to organize your data, you should ask yourself a few questions to help determine the best approach:

Date	Customer Type	Satisfaction
6/30/2012	Type 1	3.30
6/30/2012	Type 2	3.20
6/30/2012	Type 3	3.50
7/31/2012	Type 1	2.90
7/31/2012	Type 2	3.20
7/31/2012	Type 3	4.10
8/31/2012	Type 1	4.00
8/31/2012	Type 2	3.70
8/31/2012	Type 3	1.90
9/30/2012	Type 1	3.00
9/30/2012	Type 2	2.60
9/30/2012	Type 3	2.00
10/31/2012	Type 1	5.00
10/31/2012	Type 2	3.50
10/31/2012	Type 3	4.10
11/30/2012	Type 1	1.20
11/30/2012	Type 2	4.40
11/30/2012	Type 3	4.00
12/31/2012	Type 1	2.70
12/31/2012	Type 2	2.50
12/31/2012	Type 3	4.60
1/31/2013	Type 1	4.10
1/31/2013	Type 2	2.10
1/31/2013	Type 3	3.30

Figure 3. Stacked data column for satisfaction results w/ category column displaying results by customer type.

Date	Type 1 Sat	Type 2 Sat	Type 3 Sat
6/30/2012	3.3	3.2	3.5
7/31/2012	2.9	3.2	4.1
8/31/2012	4	3.7	1.9
9/30/2012	3	2.6	2
10/31/2012	5	3.5	4.1
11/30/2012	1.2	4.4	4
12/31/2012	2.7	2.5	4.6
1/31/2013	4.1	2.1	3.3

Figure 4. Multiple data columns for satisfaction results, where multiple columns are designated using customer type

Is my data column (e.g. satisfaction) defined the same way for each group?

If your data column has a different definition for each group, then you will need to present multiple data columns. For example, if satisfaction for customer type 1 was collected

differently or uses a different scale than satisfaction for customer type 2, you would need to present the data in separate columns. However, if satisfaction was collected the same and uses the same scale among all customer types, then a stacked data column approach may be the best option.

How many groups (e.g. type 1, type 2, and type 3) does my method of categorization (e.g. customer type) have?

If your data has a relatively small number of categories for which the data is displayed and only one method of categorization – organizing it as either multiple data columns or as a stacked data column w/category column would work. However, if your method of categorization has a larger number of groups – a stacked data column with category column would likely be the best method.

Is my data categorized in different ways?

As your data becomes more granular, and you introduce additional means of categorization (e.g. results are presented by date, sub-organization, and customer type) a stacked data column approach becomes the best approach for organizing your data – as the number of columns becomes too numerous.

Word of caution

Stacking data makes calculation of summary statistics, such as averages, more straightforward. However, take care when stacking survey responses, as stacking may give the perception of inflating sample size when presenting a respondent's responses as multiple records.

Should I include headings in my data?

It is not required that your file contains headings to import your data into DATAshare. However, it is a good idea to do so. Just make sure your headings (or field names) are short and meaningful to help people who are not familiar with your data better understand what it contains.

How should we order our data?

Column order (or field order) in your CSV file isn't critical for importing purposes, but does impact how your data is presented when creating visualizations, and how data is organized in raw data tables. Because of this, you should consider how your data is ordered. Below are some guidelines for you to consider when creating your dataset:

- Column order (1, 2, 3…) is based upon column (field) placement in the CSV file from left to right.

- In dropdown selections on the form to create visualizations, column order dictates which column is listed first.
- In raw data tables, column order determines how data is presented from left to right.
- Columns containing categorical data (e.g. those that will be used for grouping and filtering), should be placed to the left of numeric data used in summaries.
- If columns containing categorical data are hierarchical in nature, then the parent column should be placed to the left of the child (e.g. column containing departments, should be placed to the left of columns contain divisions, and bureaus, etc.)
- Columns containing hierarchical categorical data should be adjacent to one another.
- If numeric data columns present the same data for different time periods, older data should be placed to the left of newer data.
- If numeric data columns present sub-totals and totals for the same data, sub-totals should be placed to the left of totals.
- If not otherwise specified, columns should be ordered based on importance. More important columns should be placed to the left of less important columns.

Conclusion

Congratulations for taking the necessary steps to prepare your data for publishing. Time spent preparing your data will save you time later on. Before you proceed to the next step and publish your data, confirm that your dataset:

- ✓ Is valuable to potential users
- ✓ Contains information that is relevant to potential users
- ✓ Contains categories that will facilitate summarizing the data
- ✓ Excludes personal, confidential or sensitive data
- ✓ Is granular
- ✓ Has been saved in a CSV file format
- ✓ Includes applicable relational data
- ✓ Is stacked and ordered appropriately

Once you have confirmed the items above, publish your data by visiting DATAshare.